Review of the Banff Challenge on Upper Limits

Joel Heinrich
University of Pennsylvania

Abstract
We report the results of the Limits Challenge project, in which participants were asked to provide upper limits on a cross section s measured in a counting experiment with nuisance parameters.

1 Introduction

In July of 2006, 40 physicists and statisticians met at the Banff International Research Station (BIRS) for the *Statistical Inference Problems in High Energy Physics and Astronomy Workshop* [1] organized by James Linnemann, Louis Lyons, and Nancy Reid. Here we report on the resulting Limits Challenge project. The specification of the challenge was:

The main experiment observes events with a Poisson rate that derives from a signal of cross section s (with acceptance ϵ) and background b. Nuisance parameters (ϵ, not constrained to be ≤ 1, is actually acceptance times integrated luminosity) are measured via Poisson subsidiary measurements:

$$\begin{array}{ll} n_i \sim \mathrm{Pois}(\epsilon_i s + b_i) & \text{(main measurement)} \\ y_i \sim \mathrm{Pois}(t_i b_i) & \text{(subsidiary background measurement)} \\ z_i \sim \mathrm{Pois}(u_i \epsilon_i) & \text{(subsidiary acceptance measurement)} \end{array}$$

Channels $i = 1, 2 \ldots N$. Constants t_i and u_i are known. Upper limits (or 2-sided intervals if required by the method) on parameter of interest s to be calculated at 90% and 99% level. The $2N$ parameters ϵ_i and b_i are to be considered nuisance parameters.

It was decided that participants would provide intervals for two situations, single channel and 10 channels. The data to be used was as follows: $N = 1$: I provided a list of ~ 100000 $(n_1, y_1, z_1, t_1, u_1)$ cases for which the intervals were returned by the participants. I made coverage curves from these (using importance sampling) and calculated the Bayesian credibility of the returned intervals. $N = 10$: Same as for $N = 1$ (I provided 50 numbers for each case). Participants were warned of possible coverage problems for Bayesian methods in higher dimensions[2, 3]. The test cases I provided to the participants consisted of 3 files obtainable from [4], described as follows.

Single channel data sets: two files in ASCII text format. Each line of each file has an (n, y, z, t, u) instance for which the participants provided two upper limits: at the 90% and 99% level. (Some methods provide 2-sided intervals for some (n, y, z, t, u) combinations.) Set-1 (60229 lines) has nuisance parameters with uncertainties of about 10%, while in set-2 (39700 lines) this is increased to about 30%.

One 10 channel data set: a single file (70000 lines) in ASCII text format. Each line of each file has the (n, y, z, t, u) for each of the 10 channels (for a total of 50 numbers per line). Nuisance parameter uncertainties are about 30%. Upper limits to be provided as specified above.

2 The Submitted Methods

Eleven methods were submitted. The raw files submitted by the participants are available from [4]. Not all the participants have submitted results for all data sets. Some of the methods have built-in preferences for upper limits or 2-sided intervals. Table 1 summarizes the received entries. General reviews of strategies that have been applied to this problem are available in [5] and [6].

Table 1: Submitted methods: • for '90% and 99% intervals', ○ for '90% intervals only'.

		upper limits			2-sided intervals			
designation	type	set-1	set-2	set-3	set-1	set-2	set-3	Section
MINUIT	profile	•	•	•	•	•	•	2.1
RLC	profile′	•	•		•	•		2.1
Davison–Sartori	H-O likelihood	•	•	•				2.2
Demortier	Bayesian	•	•	•				2.3
FHC^2	mixed				•			2.4
MBT	mixed				•			2.4
Baines	Bayesian	•	•	•			•	2.5
Baines-2	Bayesian	•		•				2.5
Edlefsen	Bayesian	•	•	•				2.6
Yu	Bayesian			•				2.7
Punzi	frequentist	○	•					2.8

2.1 MINUIT and RLC

This is the profile likelihood method, submitted by Wolfgang Rolke, historically known as the MINUIT [7] method in high energy physics. Jan Conrad has written a ROOT class `TRolke` [8] that implements the scheme for Poisson upper limits in a convenient way for ROOT users. `TRolke` actually has two variations on the profile likelihood: the default 'unbounded likelihood' method (here designated 'MINUIT'), and the 'bounded likelihood' method (designated 'RLC'). The main reference for the methods is [9], which shows coverage curves that can be compared with the 1-channel coverage curves in this study. As MINUIT is based only on the likelihood, the likelihood principle is obeyed. That is, the resulting intervals depend only on the form of the likelihood, not the probability (as in the frequentist approach). Nevertheless, profile methods are neither frequentist nor Bayesian, so both the coverage and credibility are of interest in this study.

2.2 Davison–Sartori

This submission, a higher order likelihood method, is from statistics professors Anthony Davison and Nicola Sartori. The method is described in [10], which lists the following features: parameterization-invariant; computation almost as easy as first order asymptotics; more accurate than use of Bartlett correction; error $O(n^{-3/2})$ in continuous response models; gives continuous approximation to discrete response models, with error $O(n^{-1})$ at support points of the discrete distribution; relative (not absolute) error, so highly accurate in tails; see [11] for a recent review.

2.3 Demortier

This submission is from Luc Demortier. It is a Bayesian approach using reference priors for the main and subsidiary measurements considered separately, not the full reference prior, which would consider the three Poisson measurements simultaneously.

2.4 FHC^2 and MBT

Jan Conrad and Fredrik Tegenfeldt have implemented a mixed method that is Bayesian with respect to the nuisance parameters and frequentist with respect to the parameter of interest. The Poisson probability is multiplied by priors for the nuisance parameters and integrated (marginalization), leaving only a dependence on the parameter of interest. Then the unified method of Feldman and Cousins [12] is employed to extract intervals. This approach is analogous to the procedure of Cousins and Highland [13], hence the designation 'FHC^2'.

The MBT method ('modified Bayesian treatment') is a variation of the FHC^2 method described in Section 2.4, in which the ordering rule is modified. This modification is a suggestion of Gary Hill [14]. Conrad and Tegenfeldt implement MBT and compare it with FHC^2 in [15].

2.5 Baines and Baines-2

This submission is from Harvard PhD statistics student Paul Baines, who presented the matching prior approach at this conference [16]. He has provided the following brief description of the method:

The method uses a basic 'one-level' Bayesian approach (i.e. fixed hyperparameters, no hyperpriors). A limited 'grid search' was performed in a simulation study, using priors of the form:

$$p(s,b,e) \propto (s^{\alpha_s-1})(b^{\alpha_b-1})(e^{\alpha_\epsilon-1})$$

for numerous $(\alpha_s, \alpha_b, \alpha_\epsilon)$ triplets. From simulation studies, 'Pseudo-Jeffreys' priors $(1/\sqrt{\cdot})$ for the nuisance parameters and a flat prior for the interest parameter appear to perform better than most 'one-level' schemes, although slight undercoverage is expected. The approach is simple, fast to compute, and provides a benchmark for comparison with other schemes. Other 'one-level' Empirical Bayes schemes were tried with limited success. Indeed, fully Bayesian hierarchical models (e.g., as implemented by Yaming Yu) appear to offer more flexibility in accurately modelling the three-Poisson structure of the problem.

The ten-channel submission is an implementation of Jeffreys prior (i.e. $\sqrt{\det I}$) where I is the Fisher Information matrix). The one-channel entry is a minor modification of Jeffreys prior: $(1/\epsilon)*$Jeffreys. Jeffreys prior has excellent coverage properties in the absence of nuisance parameters (it is 'first order probability matching', see below). However, coverage properties are known to deteriorate in many cases when nuisance parameters are present. This implementation was used to measure the deterioration in this particular example.

His description of Baines-2 is:

This submission was another Bayesian implementation, this time using a class of priors from Tibshirani (Biometrika, 1989). I am actually giving a contributed talk at the conference about this class of priors, they are related to 'Probability Matching Priors' which give Bayesian posterior intervals with Frequentist validity. The actual submission is not of this form and is a (poor!) approximation to it. I've made some progress on this class of priors since the submission.

2.6 Edlefsen

This submission is from Harvard PhD statistics student Paul Edlefsen. He has provided the following explanation of the method:

I have produced one-sided intervals for the BIRS A1 Challenge using a numerical approximation to the Dempster-Shafer (DS) relative plausibility of singletons function. This approach results in a Bayesian posterior, but uses an intermediate calculus (DS) that is a superset of the Bayesian calculus. Unlike pure-Bayesian approaches, this does not necessitate the use of a prior. Simply put, we consider random intervals that contain the true unknown s. The intervals have distributions deduced logically from the model using the relationship between Poisson processes and exponential sums. The one-channel posterior probability distribution for s, $F(s)$, is proportional to the probability that the random interval contains s. The ten-channel distribution is proportional to the product of these one-channel distributions. The method is simpler than non-DS Bayesian methods, and requires less time to compute.

2.7 Yu

This submission is from statistics Professor Yaming Yu. He has provided the following description:

This approach treats the 10 channels as exchangeable and builds a fully Bayesian hierarchical model. We specify a common prior distribution for the nuisance parameters ϵ_i's, and vague but proper hyper-priors for the parameters of this distribution. (Likewise for the b_i's.) The hyper-priors as well as the prior on the parameter of interest (s, or source intensity) are chosen to have good frequency properties as evaluated by separate simulations. After the model is specified, posterior inference is done through Markov chain Monte Carlo. Though Monte Carlo error is present in the reported 90 and 99 percent upper bounds, it can be reduced by running a longer chain or by using more sophisticated methods to estimate quantiles from the Monte Carlo output.

2.8 Punzi

Giovanni Punzi has been developing a fully frequentist method for this problem[17, 18]. He has collaborated with Pierluigi Catastini to calculated the submitted intervals, and they provided the following summary:

The limits are obtained by implementing in a C program the method described in [17]. Limits can be produced in the same way from any desired ordering (you can have two-sided FC limits, central limits, or whatever you like), but for this challenge they were explicitly required to be upper limits. The limits are constructed to always have coverage for any value of the physical and nuisance parameters. The program takes about 1 day to run for each of the proposed files. The step size in s was 0.2, and the scan goes up to $s = 20$. This limitation has no effect on the standard coverage plots of the challenge, but causes an underestimate of the actual credibility of the intervals (we thought of this side effect only after the run).

3 Coverage

Frequentist coverage is the first criterion by which the submissions are compared. The coverage probability is defined in the single channel case as

$$C(s, b, \epsilon) = \sum_{}^{'} \frac{e^{-\mu}\mu^n}{n!} \frac{e^{-\nu}\nu^y}{y!} \frac{e^{-\rho}\rho^z}{z!}$$

where $\mu = \epsilon s + b$, $\nu = tb$, and $\rho = u\epsilon$, and ϵ and b are fixed representative values for the coverage calculation (t and u are fixed values specified without uncertainty). Here $\sum'$ means sum only over values of (n, y, z) that yielded an interval that includes s.

This is the classic definition of frequentist coverage probability. s, ϵ and b are thought of as the 'true' values of the parameters that are unknown in real life. One investigates how the method performs for (representative) fixed true values of the parameters.

The 'true' values somewhat arbitrarily selected for the nuisance parameters to produce $C(s)$ for $0 \leq s \leq 20$ are:

set-1: $b = 3$, $\epsilon = 1$

set-2: $b = 3$, $\epsilon = 1$

set-3: $b_i = 0.31$, $\epsilon_i = 0.1$

Because of the range of (n, y, z, t, u) values provided in the 3 sets, the b and ϵ values assumed for a plot of $C(s)$ can be varied somewhat, e.g., for set-1 $\sim 2.5 \leq b \leq\sim 3.5$ is doable, but not much further outside that range. But no significant changes were observed for other values in the allowable range, so just one representative set is shown here.

Figure 1 shows $C(s)$ for selected 90% intervals, and Fig. 2 shows 99% intervals. Coverage curves for all submitted sets are available at [4]. Briefly summarizing:

– MINUIT covers at $\sim$ nominal for sets 1 and 2; set-3 is a bit lower, but is still acceptable.

– RLC's coverage can oscillate in the $0 < s < 5$ region, but otherwise OK.
– Davison–Sartori often undercovers at small s.
– Demortier undercovers in set-3.
– FHC^2 and MBT overcover slightly.
– Baines undercovers set 3.
– Baines-2 shows slight undercoverage.
– Edlefsen covers $\sim$ nominal for sets 1 and 3; for set-2 overcovers.
– Yu shows slight undercoverage.
– Punzi shows moderate overcoverage.

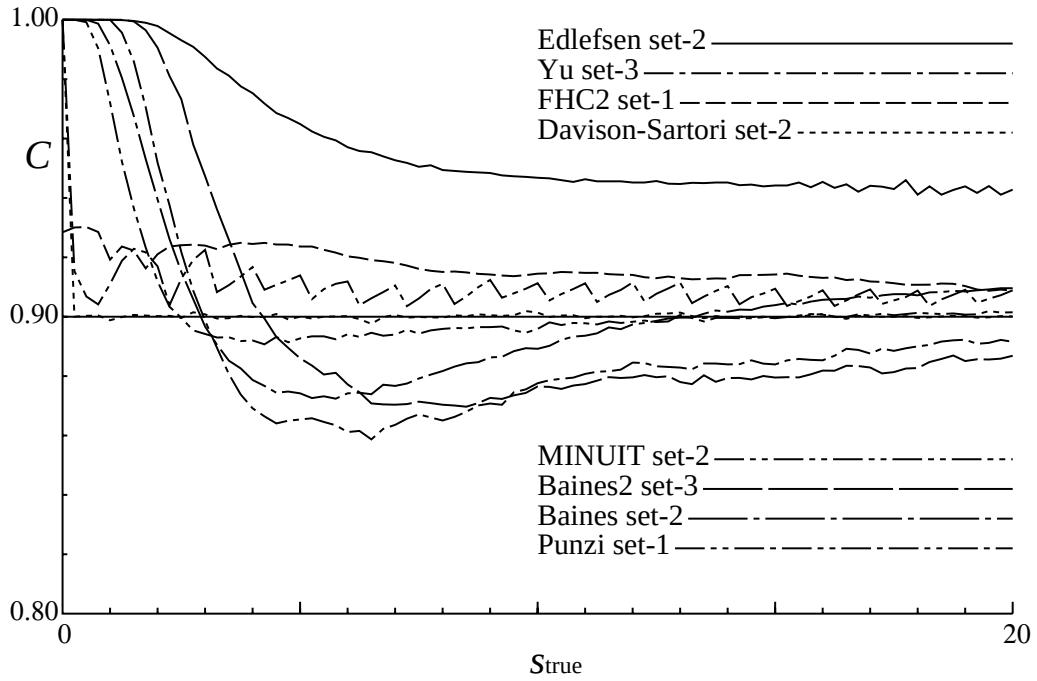

Fig. 1: Coverage of selected 90% intervals as a function of the true value of s.

All methods with submitted results on all 3 data sets show at least moderate deviations from nominal coverage (either overcoverage or undercoverage) for at least one of the sets, but most of these methods still seem usable. MINUIT, for example, achieves coverage properties similar to the more sophisticated Bayesian methods, but with much less computational cost.

4 Credibility

To further characterize the performance, one would like to have the Bayesian credibility for each of the supplied intervals. While the coverage calculation is completely specified by the definition, calculating the Bayesian credibility of the intervals supplied by the participants presents a bit of a problem, as one needs to select priors for the parameter of interest and the nuisance parameters. I have somewhat arbitrarily selected the following priors:

sets 1 and 2: flat for s, b and ϵ

set-3: flat for s. Priors for b and ϵ are $b_i^{-0.9}$ and $\epsilon_i^{-0.9}$

The priors for the nuisance parameters (applied to the likelihood for the auxiliary measurements) in the 10-channel case are chosen so that the effective priors for the total background $b' = \sum_i b_i$ and total acceptance $\epsilon' = \sum_i \epsilon_i$ are flat.

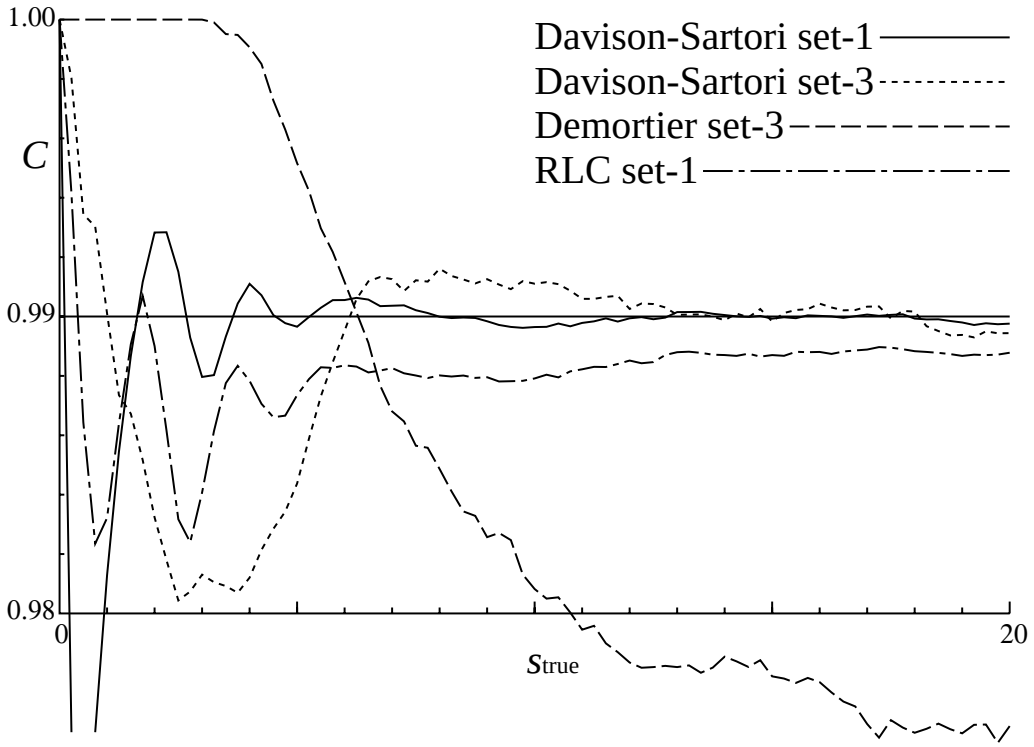

Fig. 2: Coverage of selected 99% intervals as a function of the true value of s.

Sample distributions of credibilities are shown in Figs. 3–5; see [4] for a complete set.

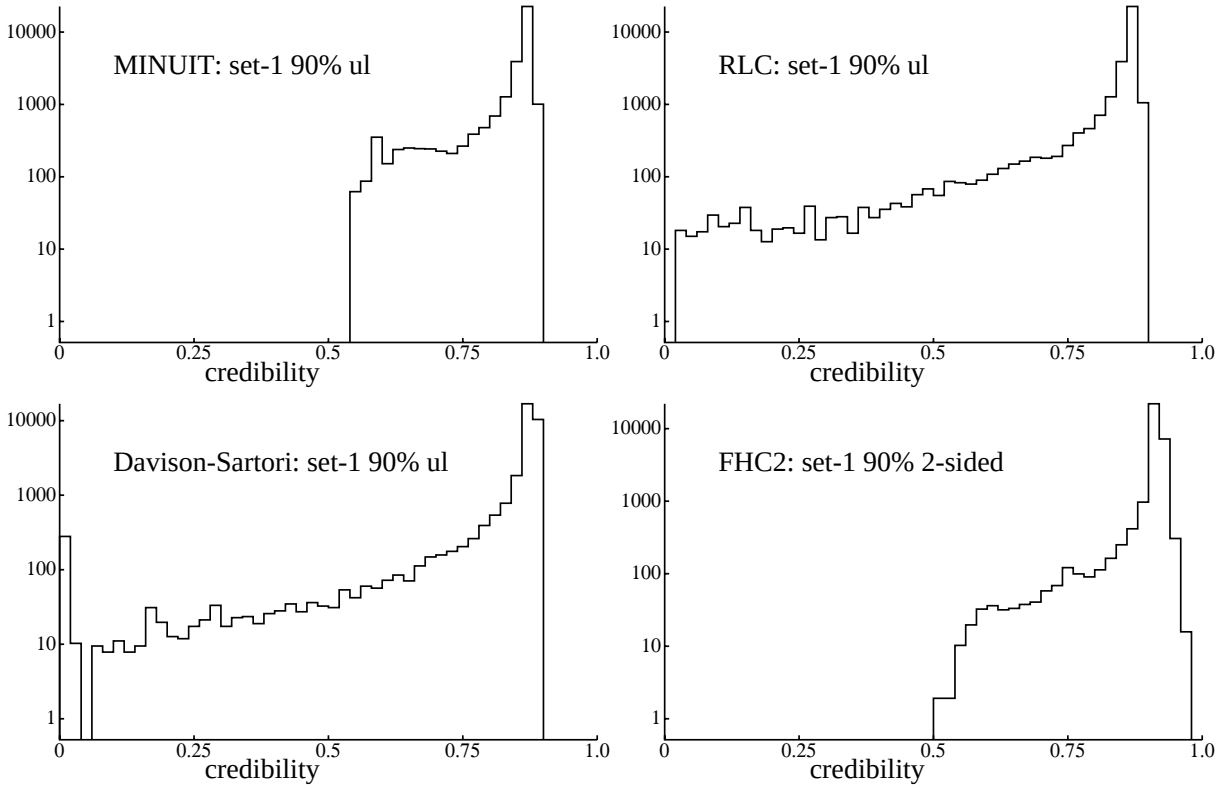

Fig. 3: Distribution of set-1 credibilities for selected methods.

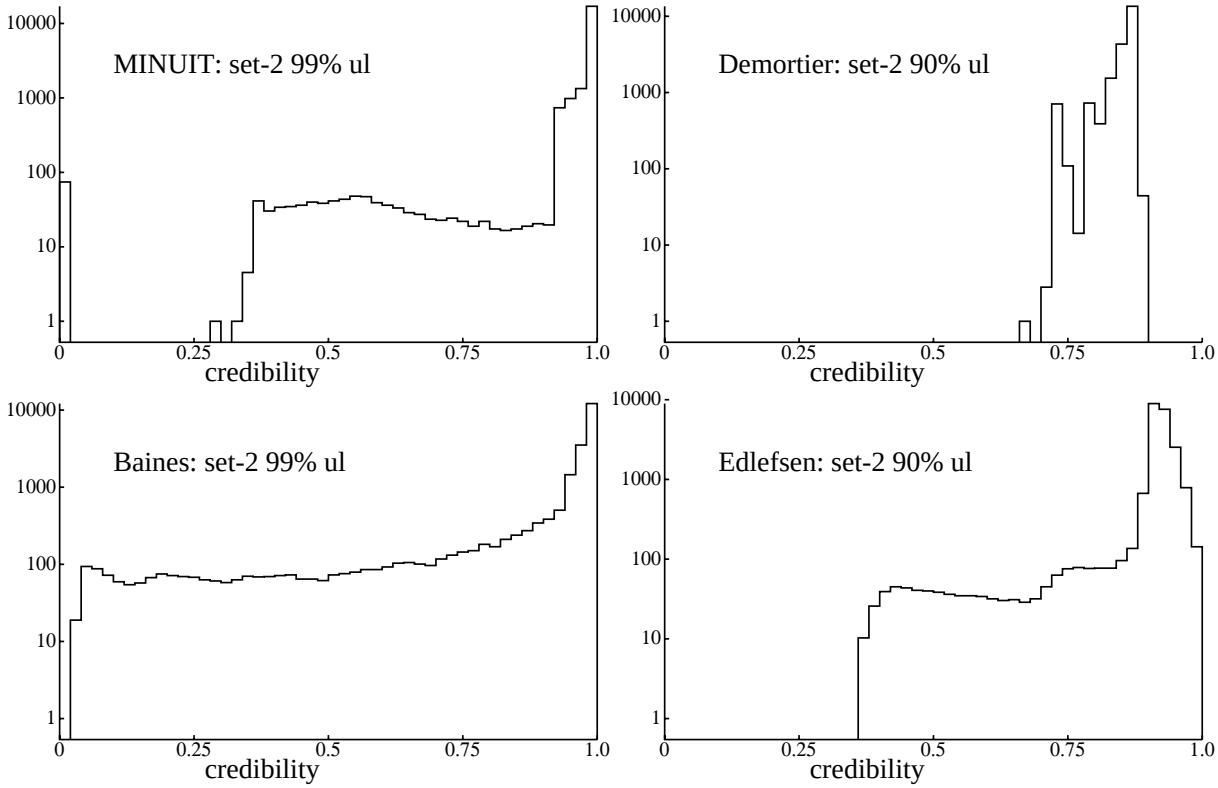

Fig. 4: Distribution of set-2 credibilities for selected methods.

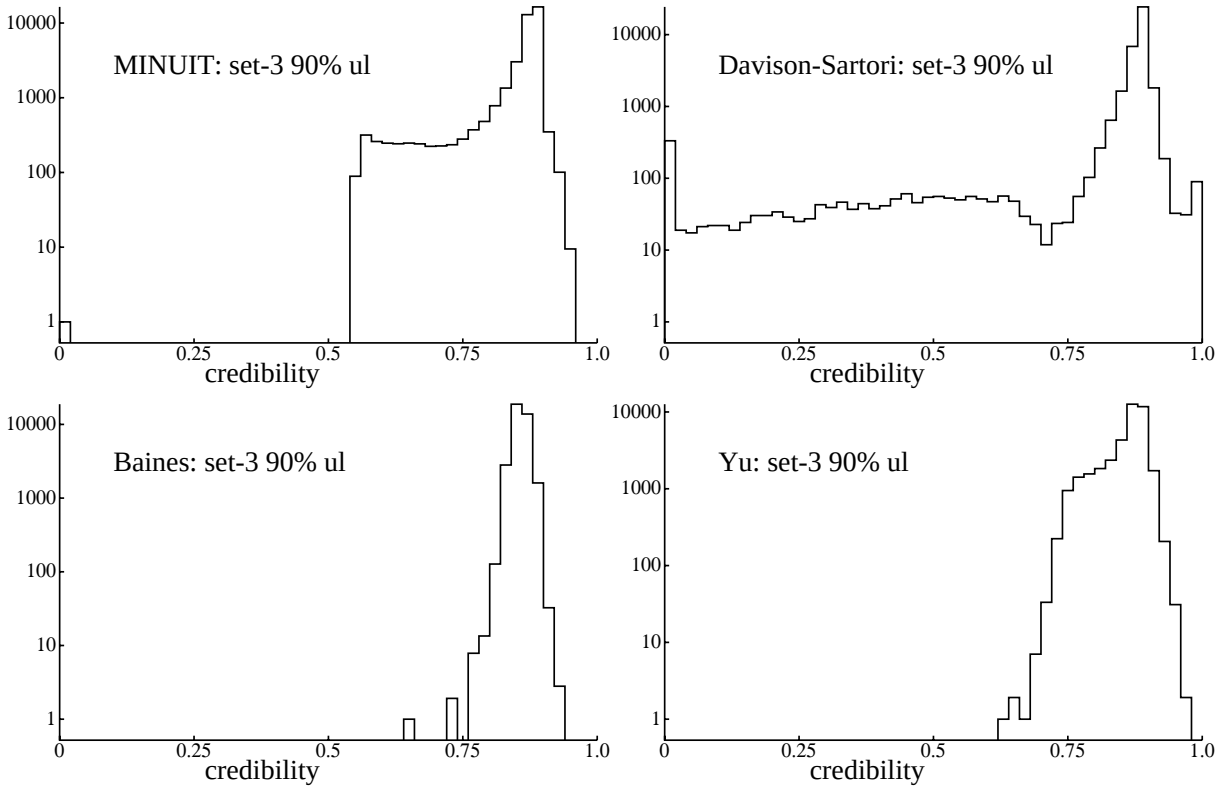

Fig. 5: Distribution of set-3 credibilities for selected methods.

Large deviations in credibility from nominal require investigation, but as the choice of prior is not unique, moderate deviations are not considered significant.

Table 2 shows some set-1 upper limits and calculated credibilities for comparison. The estimate of b (based on the observed y) increases as one moves down the table. RLC shows some intervals with rather low credibility. Focusing on the set-1 90% upper limits, one finds some intervals with credibilities as low as 2%. With $n = 1$, as y, the number of background events observed in the subsidiary background measurement, increases, the resulting upper limit drops rapidly to 0.02 at $y = 95$, then jumps discontinuously to 1.4 at $y = 96$:

Table 2: Selected 90% upper limits with credibilities for set-1 with $n = 1$, $z = 111$, $t = 33$, $u = 100$.

	RLC		D–S		Punzi		FHC^2		MINUIT	
y	ul	cred	ul	cred	ul	cred	ul	cred	ul	cred
79	0.458	0.307	0.727	0.445	1.0	0.560	2.013	0.819	1.13	0.606
84	0.322	0.229	0.591	0.383	0.8	0.483	1.861	0.796	1.10	0.601
89	0.187	0.141	0.455	0.313	0.6	0.392	1.664	0.760	1.08	0.598
90	0.160	0.122	0.428	0.297	0.6	0.393	1.664	0.760	1.08	0.599
91	0.133	0.103	0.401	0.282	0.6	0.393	1.664	0.761	1.07	0.599
92	0.106	0.083	0.374	0.266	0.6	0.394	1.664	0.761	1.07	0.597
95	0.025	0.020	0.292	0.216	0.4	0.284	1.664	0.763	1.06	0.595
96	1.366	0.692	0.265	0.198	0.4	0.284	1.664	0.764	1.05	0.592
98	1.312	0.678	0.211	0.162	0.4	0.285	1.512	0.731	1.05	0.594
101	1.230	0.656	0.130	0.103	0.4	0.287	1.512	0.732	1.04	0.592
103	1.175	0.640	0.076	0.061	0.2	0.155	1.512	0.734	1.03	0.590
107	1.066	0.605	0.000	0.000	0.2	0.156	1.375	0.701	1.02	0.589
114	0.876	0.537	0.000	0.000	0.0	0.000	1.375	0.704	1.00	0.586

Davison–Sartori also shows intervals with low credibility. As the background estimate increases, the upper limit drops gradually to zero, and stays there. Punzi, implementing a fully frequentist method, shows similar behaviour.

One of the benefits of the unified method of Feldman and Cousins is that it tends to avoid this behaviour. MINUIT also shows good performance with respect to this criterion; the credibility staying reasonably large:

4.1 Behaviour with zero observed events

With $n = 0$, the Poisson likelihood is $\exp[-(\epsilon s + b)]$. The shape of the likelihood with respect to the parameter of interest s is, in this special case, independent of the true value of b. Methods that obey the likelihood principle will consequently show no dependence of the upper limit for s on the background estimate or its uncertainty.

Alternatively: When zero events are observed in the main measurement, one knows that *zero signal* events were observed (and also zero background events). For the $n = 0$ special case, we have absolute separation between signal and background; consequently the uncertainty associated with not knowing if the events were signal or background is absent.

I check each submitted method to see whether the resulting intervals depend on the background estimate. Looking at set-1, for example, MINUIT, Demortier, Baines, Baines-2, and Edlefsen demonstrate background-independent $n = 0$ intervals.

For set-1, both Davison–Sartori and Punzi always produce zero-length 90% intervals whenever $n = 0$. As shown in Table 3, RLC and FHC^2 show a strange dependence of the limit on the background estimate when $n = 0$, and at 99%, Davison–Sartori shows a few rather narrow but finite intervals.

Table 3: Selected upper limits for set-1 with $n = 0$, $z = 110$, $t = 33$, $u = 100$.

y	RLC 90%	FHC^2 90%	D–S 99%
84	0.325	0.908	0.180
90	0.161	0.825	0.017
102	1.213	1.000	0.000
112	0.939	0.908	0.000
119	0.746	0.825	0.000
128	0.500	0.750	0.000

5 Conclusions

Comparison of the submitted results leads to the following conclusions about the performance of the methods. The conclusions are specific to the particular type of Poisson problem investigated in this project; they will not necessarily generalize to other applications (measurements of particle masses or lifetimes, for example) or to 5-σ confidence level. Comparing the methods:

- Overall, MINUIT (i.e. profile) is the easiest of the methods computationally, and its performance seems quite acceptable on the whole. The RLC variant performs less well, and was not provided for the 10-channel case.
- The fully Bayesian methods can perform excellently, but take more computational effort. One needs some care in selecting the priors, especially for the 10-channel case.
- The FHC^2 and MBT methods (mixed frequentist-Bayesian providing two sided intervals) behave well in general with respect to the coverage and credibility criteria, but it's not numerically clear what happens when $n = 0$ events are observed. (Of course, the frequentist component of FHC^2 and MBT does not necessarily satisfy the likelihood principle.)
- The fully frequentist method of Punzi and the higher order likelihood method of Davison–Sartori can produce zero-length or excessively narrow (i.e. low credibility) intervals. Punzi is not yet available for 10-channels. Davison–Sartori shows oscillations of coverage.

General conclusions are:

- Bugs are a ubiquitous problem; no software package is immune. Coverage and credibility checks were useful in uncovering some of these bugs. (Several of the entries were re-submitted after the initial coverage plots were viewed by the submitters.)
- Coverage is a well defined performance criterion. Bayesian credibility depends on the choice of prior(s), but intervals with very low credibility are worth investigating.
- Zero-length intervals are widely viewed as undesirable; very low credibility intervals seem undesirable for essentially the same reasons. Nevertheless, a document *Why Frequentists Should Care About Bayesian Credibility* may be necessary to convince hard core frequentists. (Does such a document already exist?)
- The companion document *Why Bayesians Should Care About Frequentist Coverage* would also be useful, and probably already exists.
- The Limits Challenge project has attracted significant interest, including both physicists and statisticians. It seems likely that after the PHYSTAT-LHC workshop more submissions will be sent to fill some of the gaps (or to fix some bugs) still present in the current submissions. These are certainly welcome.
- It would be useful to preserve the software that calculates the coverage and credibility, as well as the data sets and submitted files.

Acknowledgements

I would like to thank the organizers of the PHYSTATLHC and BIRS Conferences, and the many statisticians and physicists who submitted their methods to this project.

References

[1] http://www.pims.math.ca/birs/birspages.php?task=displayevent\&event_id=06w5054

[2] J. Heinrich, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press, 98 (2006)

[3] http://www.samsi.info/200506/astro/workinggroup/phy/jgh.pdf

[4] http://newton.hep.upenn.edu/~heinrich/birs/

[5] R. Cousins, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press, 75 (2006)

[6] J. Heinrich and L. Lyons, "Systematic Errors", in *Annual Review of Nuclear and Particle Science*, Vol. 57, Palo Alto, Annual Reviews, 145 (2007)

[7] F. James, *MINUIT—Function Minimization and Error Analysis*, Version 94.1, CERN Program Library Long Writeup D506, (1994)

[8] http://root.cern.ch/root/html/TRolke.html

[9] W. Rolke, A. Lopez, and J. Conrad, Nucl. Instrum. Methods Phys. Res. A 551/2-3, 493 (2005)

[10] http://www.imsv.unibe.ch/htm/sstats/NicolaSartori.pdf
http://newton.hep.upenn.edu/~heinrich/birs/davison-sartori/doc/

[11] A.R. Brazzale, A.C. Davison and N. Reid, *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge, Cambridge University Press (2007)

[12] G.J. Feldman and R.D. Cousins, Physical Review D **57**, 3873 (1998)

[13] R.D. Cousins and V.L. Highland, Nucl. Instrum. Methods Phys. Res. A 331 (1992)

[14] Gary C. Hill, Physical Review **D67**, 118101 (2003)

[15] J. Conrad and F. Tegenfeldt, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press 93 (2006)

[16] Paul Baines, *Probability matching priors in LHC Physics—a pragmatic approach*, these proceedings

[17] G. Punzi, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U. M. Karagoz, eds. London: Imp. Coll. Press 88 (2006)

[18] http://www.samsi.info/200506/astro/workinggroup/phy/SAMSI_punzi.ppt.pdf